Condition and on-site Sampling of Count Data Models

Conditional and on-site sampling of count data models

Maki Momma

Contents

1. Introduction

It is often assumed in theoretical studies that statistical inference is based on data chosen randomly from the entire population. This is not, however, a realistic assumption in many cases, and often we are forced to work with biased or incomplete data. In economics and other social sciences, where controlled experiments are difficult, this tendency is even more prominent. When data are obtained by a survey for example, individuals have the freedom to choose whether or not to participate in the study, and this is likely to cause bias in the sample. In other cases, we choose to collect data from a sub-population, not the entire population, and this could also result in a biased sample. Two methods of collecting data from a sub-population are the so-called conditional sampling and on-site sampling.

Conditional sampling refers to the sampling scheme where data are observed under the condition their values exceed a certain number, most likely zero. An example is when a study is

conducted on the number of traffic violations, and data are collected from the list of those who have a previous record of traffic violations. By nature, this method excludes zero values in the sample. Another sampling scheme often confused with conditional sampling is on-site sampling. When we want to study how the number of visits to the doctor depend on the patient's age, sex, etc., we often choose to take a sample from patients who are present at a hospital (on-site) on a particular date and at a particular time. A sample obtained in this fashion not only excludes zero values but is also biased toward larger values. Nevertheless, this type of sampling is employed regularly since it is often easier to take samples from items known to take positive values, especially when many zeros are expected in the data. In addition, if used properly, an on-site sample is likely to produce more accurate parameter estimates than a random sample based on the same sample size. The important point is not so much in trying to avoid biased sampling schemes but to acknowledge the type and magnitude of the bias so as to correctly assess its relation to the parent population.

This paper provides information on the bias of the two sampling methods described above, namely, conditional sampling and on-site sampling In addition, this study describes the characteristics of each method and discusses their relations to each other and to the distribution of the parent population. The two methods are often confused with one another in empirical studies, but the distributional features are distinct and ignoring their differences could result in a misleading conclusion.

Discussion is limited to the sampling of discrete variables, since conditional sampling method is valid only for count data, and the emphasis is on cases where a count variable is dependent on quantitative explanatory variables, known as count data models. Note that in each of the two sampling schemes considered, the method itself involves random sampling, but the population from which the sample is taken is not the entire population, but a biased sub-population, and hence results in a biased sample.

In Section 2, conditional sampling and on-site sampling methods are reviewed for discrete variables in general. Both methods produce biased samples of the parent population, and among these two, on-site sampling method has larger bias as well as larger variance. Section 3 deals with count data models and distributions of response variables under conditional and on-site sampling methods. Section 4 extends these models to cases where data are also truncated from above, and likelihood functions are then derived for both sampling methods. Section 5 concludes this research. Throughout this paper, $p(y) = P(Y = y)$ denotes the probability of a discrete random variable Y

in the parent population, unless otherwise noted.

2. Truncated at zero models--- biased sampling

2-1 Conditional sampling

When we want data on the number of times an individual participates in a particular activity or event, it is sometimes easier to take a sample from the list of those who have participated in the activity (event) at least once. If this type of sampling is employed, individuals without any previous record of participation have no possibility of being sampled. In other words, the obtained data will be truncated at zero. This type of sampling will be referred to as conditional sampling. For example, when a list is generated from the E-mail addresses of people who visit a paritcular website and a sample is chosen randomly from that list, this would be conditional sampling.

Under this sampling scheme, at least one occurrence of the event of interest,i.e.participation in a particular activity is the necessary condition to be included in the list, which will be called the conditional population. A conditional population, then, is a sub-population composed of individuals taking positive values only. The probability of the number of event occurrences for individuals in the conditional population is the probability of the count variable conditioned to take positive values, and so is given by

$$p_C(y) = p(y \mid Y > 0) = \frac{p(y)}{P(Y > 0)} = \frac{p(y)}{1 - p(0)}, \tag{2.1}$$

where the subscript C stands for conditional population, and $p(y) = p(Y = y)$ denotes the probability function of the count variable Y in the parent population. The expected value and variance of the conditional population are derived as

$$E(Y_C) = \frac{E(Y)}{P(Y > 0)} = \frac{\sum_{k=1}^{\infty} P(Y > k)}{P(Y > 0)} \tag{2.2}$$

and

$$V(Y_C) = \frac{1}{P(Y>0)}\left(\sum_{y=1}^{\infty} y^2 p(y) - \frac{\{E(Y)\}^2}{P(Y > 0)} \right) \tag{2.3}$$

respectively. Here, $E(Y)$ denotes the expected value of the count variable in the parent population. When $p(0) = P(Y > 0) > 0$, the expected value of the conditional population exceeds that of the parent population. In other words, this sampling procedure is biased. It can also be shown in a straightforward manner that $V(Y_C) < V(Y)$,regardless of the distributional form of Y ,

that the variance of the population of individuals conditioned to take positive values is smaller than that of the parent population.

As an illustrative example, let us consider the case where the count variable follows a Poisson distribution. The Poisson model is the most basic model of discrete variables, and is given by

$$p(y) = \frac{e^{-\lambda}\lambda^{y}}{y!}, \tag{2.4}$$

where λ is the mean of the parent population. Probability of the corresponding count variable Y_C in the conditional population is given by

$$p_C(y) = p(y|y>0) = \frac{\dfrac{e^{-\lambda}\lambda^{y}}{y!}}{\displaystyle\sum_{i=1}^{\infty}\frac{e^{-\lambda}\lambda^{i}}{i!}} = \frac{\dfrac{e^{-\lambda}\lambda^{y}}{y!}}{1-e^{-\lambda}} = \frac{e^{-\lambda}\lambda^{y}}{y!(1-e^{-\lambda})},$$

from which the expected value is obtained as

$$E(Y_C) = \frac{\lambda}{1-e^{-\lambda}} = 1 + \frac{\lambda}{2} + \sum_{k=1}^{\infty} B_{2k}(-\lambda)^{2k} \qquad \text{where } B_{2k} \text{ is a Bernoulli number.}$$

Variance is given by

$$V(Y_C) = \frac{\lambda}{1-e^{-\lambda}} - \frac{\lambda^2 e^{-\lambda}}{\left(1-e^{-\lambda}\right)^2}.$$

Distribution of the conditional population is clearly under-dispersed, that is, the mean exceeds the variance. Note that mean and variance in the parent population are equal (referred to as equi-dispersion) for the Poisson model.

2-2 On-site sampling

When searching for oil, larger oil fields are more likely to be discovered than smaller ones because of their sizes. This type of occurrence, well acknowledged for cases with continuous variables, is known as length-biased or size-biased sampling. Since larger values are likely to be included in the data, the obtained sample is biased and an estimation should be carried out with this in mind. What is less known is the fact that similar circumstances exist when sampling discrete data. When we want data on the number of visits to a certain site and we choose an on-site sample, i.e. a sample chosen randomly from the population of individuals present at the site of interest on a pre-selected date and time (the on-site population), then individuals will have a

bigger chance of being selected if they are often at the sampling site. This point is overlooked in many empirical studies, and often, conditional distribution of the previous section is used to estimate when in fact it is more appropriate to use the on-site distribution as derived below.

While on-site sampling method requires choosing items randomly on-site, it not only fails to pick up zero values but also produces positive bias in the sample. As already stated, this stems from the fact that on-site population is a sub-population composed of individuals on-site at the time of sampling. Individuals often at the site have a higher probability of being in the on-site population, and thus a higher probability of being selected in the sample. Probability of a count variable in an on-site population is related to that of the parent population by

$$p_B(y) = \frac{yp(y)}{E(Y)} = \frac{yp(y)}{\sum_{k=0}^{\infty} P(Y > k)}, \tag{2.5}$$

under the assumption that the probability of being on-site is proportional to the value it takes. Note that it is more accurate, in this case, to assume that an item's probability of being on-site is proportional to the length of time spent at the sampling site. This, however, causes significant complexity in the model, whereas results obtained using such models are most likely, barely different from the simplified version employed here. The mean and variance of a count variable Y_B in an on-site population are related to that of the parent population by

$$E(Y_B) = \frac{E(Y^2)}{E(Y)} \tag{2.6}$$

and

$$V(Y_B) = \frac{1}{E(Y)}\left(E(Y^3) - \frac{\{E(Y^2)\}^2}{E(Y)} \right) \tag{2.7}$$

respectively. The expected value of Y_B exceeds the mean of the parent population, since $E(Y_B) - E(Y) = \frac{1}{E(Y)}\left[E(Y^2) - \{E(Y)\}^2\right]$. Variance of Y_B exceeds that of the parent population if and only if

$$E(Y)\{E(Y^3) + [E(Y)]^3\} > E(Y^2)\{E(Y^2) + [E(Y)]^2\}.$$

For Poisson distributed count variables, $E(Y_B) = \frac{\lambda(\lambda+1)}{\lambda} = \lambda + 1$ and $V(Y_B) = \lambda$, so the distribution is also under-dispersed using this sampling scheme.

2-3 Relation between conditional and on-site sampling

Intuitively, it seems that bias of a count variable in an on-site population should be larger in

magnitude than bias in the conditional population, since on-site sampling has a tendency to select larger values. A formal proof is given below:

Theorem For Y discrete, $E(Y_B) \geq E(Y_C)$ regardless of the form of the distribution.

Proof. It is necessary and sufficient to show that

$\frac{E(Y^2)}{E(Y)} \geq \frac{E(Y)}{1-p(0)}$ regardless of the form of the distribution. This is equivalent to showing that

$E(Y^2)\{P(Y>0)\} \geq \{E(Y)\}^2$, which is a sharper version of the well known Cauchy-Schwartz inequality. But,

$$
\begin{aligned}
& E(Y^2)\{P(Y>0)\} - \{E(Y)\}^2 \\
&= \sum_{y=1}^{\infty} y^2 p(y) \sum_{y=1}^{\infty} p(y) - \left\{ \sum_{y=1}^{\infty} y p(y) \right\}^2 \\
&= \sum_{k=1}^{\infty} k^2 p(k) + \sum_{j<k} (j^2 + k^2) p(j) p(k) - \sum_{k=1}^{\infty} k^2 p(k) - 2 \sum_{j<k} jk p(j) p(k) \\
&= \sum_{j<k} (j-k)^2 p(j) p(k) \\
&> 0.
\end{aligned}
$$

Similarly, it can be shown that variance of a count variable in an on-site population exceeds that of the conditional population.

Lemma For Y discrete, $V(Y_B) \geq V(Y_C)$ regardless of the form of the distribution.

Proof. It is necessary and sufficient to show that

$$
\frac{1}{E(Y)} \left(E(Y^3) - \frac{\{E(Y^2)\}^2}{E(Y)} \right) > \frac{1}{P(Y>0)} \left(\sum_{y=1}^{\infty} y^2 p(y) - \frac{\{E(Y)\}^2}{P(Y>0)} \right)
$$

which is equivalent to showing that

$$
\{P(Y>0)\}^2 \left(E(Y^3) E(Y) - \{E(Y^2)\}^2 \right) > \{E(Y)\}^2 P(Y>0) \sum_{y=1}^{\infty} y^2 p(y) - (E(Y))^4 .
$$

To prove the inequality of above, it is sufficient to show that

$$
\{E(Y)\}^4 + E(Y^3) E(Y) \{P(Y>0)\}^2 - \{P(Y>0)\}^2 \{E(Y^2)\}^2 - \{P(Y>0)\} \{E(Y^2)\} \{E(Y)\}^2 > 0 .
$$

Rearranging terms, it can be shown, as with the case of the mean that

$$
\begin{aligned}
&\{E(Y)\}^4 + E(Y^3)E(Y)\{P(Y>0)\}^2 - \{P(Y>0)\}^2\{E(Y^2)\}^2 - \{P(Y>0)\}\{E(Y^2)\}\{E(Y)\}^2 \\
&= \left[\{E(Y)\}^4 - \{P(Y>0)\}^2\{E(Y^2)\}^2\right] + \left[E(Y^3)E(Y)\{P(Y>0)\} - \{P(Y>0)\}\{E(Y^2)\}\{E(Y)\}^2\right] \\
&> \left[E(Y^3)E(Y)\{P(Y>0)\} - \{P(Y>0)\}\{E(Y^2)\}\{E(Y)\}^2\right] \\
&= \sum_{j<k}(k^2 - j^2)(k-j)p(j)p(k) \\
&> 0.
\end{aligned}
$$

To study the distributional differences of the two sampling schemes in more detail, recall that when the distribution of a count variable Y in the parent population is Poisson, the corresponding distribution of Y_C in a conditional population is given by $p_C(Y) = \frac{1}{1-e^{-\lambda}}\frac{e^{-\lambda}\lambda^y}{y!}$, whereas, distribution of Y_B in an on-site population is given by $p_B(Y) = \frac{e^{-\lambda}\lambda^{y-1}}{(y-1)!}$. For Y_C, probability $p_C(k)$ for every k is simply the probability of the parent population multiplied by a constant $1-e^{-\lambda}$. The relative magnitudes of the probabilities, therefore, are identical to that of the parent population. On the other hand, the probability distribution of Y_B is displaced (shifted) entirely to the right. This distinction occurs since on-site sampling has a tendency to choose variables with larger values, regardless of the distributional form of Y. Ignoring these fundamental differences and confusing the two methods results in an incorrect assessment of the parent population. It can indeed be shown that for any discrete Y, probability of Y_B exceeds that of Y_C for values k such that $k > \frac{E(Y)}{1-p(0)}$.

3. Count data models

3-1 Multiplicative heterogeneity models

In many cases, the objective of studying discrete data is to specify the dependence of a count variable on one or more quantitative or qualitative observable variables. A typical approach to this problem is to employ regression techniques. The most basic model of count regression is the simple Poisson model, where the response count variable Y follows a Poisson distribution

$$P(Y=y|\lambda) = p(y|\lambda) = \frac{e^{-\lambda}\lambda^y}{y!},$$

and the expected value $\lambda = E(Y)$ is related to the m dimensional column vector of regressors $x = (x_1,\ldots,x_m)'$ by the relation $\log\lambda = x'\beta$. Note that β is a $m\times 1$ parameter vector $(\beta_1,\ldots,\beta_m)'$ to be estimated from the data. Then, $E(Y) = \lambda = e^{x'\beta}$, so the specification of above ensures positivity of the expected value of the counts.

The Poisson model does not allow heterogeneity in the population and also imposes a restriction that the variable is equi-dispersed. Its use in practice, therefore, is limited. To allow heterogeneity, it is often assumed that $E(Y_i) = \widetilde{\lambda}_i = \lambda_i \nu_i = \exp(x_i'\beta)\nu_i$ for every observation i, where ν_i is an unobservable random variable representing heterogeneity of item i. This type of model is called a multiplicative heterogeneity model. For identification purpose, ν is standardized to have expected value one, i.e., $E(\nu) = 1$. Since ν cannot be observed, we need to integrate this factor out to estimate the model. Letting g denote the density of ν, the marginal probability of count variable Y with multiplicative heterogeneity given the values of explanatory variables, is derived as

$$p(y \mid x) = \int \frac{e^{-\lambda\nu}(\lambda\nu)^y}{y!} g(\nu) d\nu , \tag{3.1}$$

a mixed Poisson distribution.

Probability of Y_C in a conditional population, for this case, is given, using (2.1) and (3.1),by

$$p_C(y|x) = \frac{p(Y)}{P(Y>0)} = \frac{\int \frac{e^{-\lambda\nu}(\lambda\nu)^y}{y!} g(\nu) d\nu}{1-\varphi(\lambda\nu)}, \tag{3.2}$$

from which the expected value and variance are derived as

$$E(Y_C) = \frac{E(Y)}{P(Y>0)} = \frac{\lambda E(\nu)}{1-\varphi(\lambda\nu)} = \frac{\lambda}{1-\varphi(\lambda\nu)} \tag{3.3}$$

and

$$V(Y_C) = \frac{1}{1-\varphi(\lambda\nu)}\left(\lambda^2 E(\nu^2) + \lambda E(\nu) - \frac{\lambda^2\{E(\nu)\}^2}{1-\varphi(\lambda\nu)}\right) = \frac{1}{1-\varphi(\lambda\nu)}\left(\lambda^2 E(\nu) + \lambda - \frac{\lambda^2}{1-\varphi(\lambda\nu)}\right), \tag{3.4}$$

respectively. Note here that φ denotes the characteristic function of the heterogeneity factor ν. This distribution is under-dispersed if $1-\varphi(\lambda\nu) < \frac{\{E(\nu)\}^2}{E(\nu^2)}$ or equivalently, if $\{1-\varphi(\lambda\nu)\}E(\nu^2) < 1$,and over-dispersed if $\{1-\varphi(\lambda\nu)\}E(\nu^2) > 1$.

For on-site population, probability of Y_B is given, using (2.5) and (3.1), by

$$p_B(y|x) = \frac{yp(y|x)}{E(Y)} = \frac{\int \frac{e^{-\lambda\nu}(\lambda\nu)^y}{(y-1)!} g(\nu) d\nu}{\lambda}, \tag{3.5}$$

from which the expected value and variance are derived as

$$E(Y_B) = 1 + \lambda E(\nu^2) \tag{3.6}$$

and

$$V(Y_B) = \lambda E(v^2) + \lambda^2 E(v^3) - \lambda^2 \{E(v^2)\}^2 \tag{3.7}$$

respectively. Over-dispersion occurs if and only if $Ev^3 > \{E(v)^2\}^2 + \frac{1}{\lambda^2}$

3-2 A parametric example

Depending on the parametric form of the density g, various heterogeneous count data models can be proposed. A popular method is to assume that heterogeneity factor v follows a Gamma distribution $\Gamma(\alpha,\alpha)$, where the value of parameter α is to be estimated from the data. For specification purpose, the distribution is standardized to have expected value one. This assumption leads to a Negative Binomial distribution for the count variable Y, that is

$$p(y \mid x,\lambda,\alpha) = \frac{\Gamma(\alpha+y)}{\Gamma(\alpha)\Gamma(y+1)}\left(\frac{\alpha}{\lambda+\alpha}\right)^{\alpha}\left(\frac{\lambda}{\lambda+\alpha}\right)^{y}. \tag{3.8}$$

Many empirical studies show that negative binomial model does indeed fit the data very well, and therefore may be considered as a preliminary candidate of a parametric count data model with multiplicative heterogeneity. Expected value and variance of Y for this model are given by λ and $\lambda + \frac{\lambda^2}{\alpha}$. Using (3.3), (3.4) and (3.8), the mean and variance of Y_C in a conditional population are derived as

$$E(Y_C) = \frac{\lambda}{1-\left(\frac{\alpha}{\alpha+\lambda}\right)^{\alpha}}$$

and

$$V(Y_C) = \frac{1}{1-\left(\frac{\alpha}{\alpha+\lambda}\right)^{\alpha}}\left\{\lambda^2\left(1+\frac{1}{\alpha}-\frac{1}{1-\left(\frac{\alpha}{\alpha+\lambda}\right)^{\alpha}}\right)+\lambda\right\}$$

respectively. The distribution is over-dispersed if and only if $\left(\frac{\alpha}{\alpha+\lambda}\right)^{\alpha} < \frac{1}{\alpha+1}$. Mean and variance of Y_B in an on-site population are given, using (3.6) and (3.7) by

$$E(Y_B) = \frac{\alpha+\lambda+\alpha\lambda}{\alpha} = 1+\lambda+\frac{\lambda}{\alpha}$$

and

$$V(Y_B) = \frac{\lambda(\alpha+1)(\alpha+\lambda)}{\alpha^2}$$

respectively. Over dispersion occurs when $\lambda > \frac{\alpha^2}{\alpha+1}$.

4. Right truncated models

When making observations on a discrete variable, data are often truncated from above. A typical case is when data are collected in surveys, and participants are asked to count the number of certain events, with the highest category in the questionnaire being "n or more". In this section, estimation methods of count data models with this type of censoring, where observations are truncated from above in a deterministic manner, are discussed. It is assumed that sampling is done either from a conditional or an on-site population. In other words, the observed data are censored both from the right above and the left below.

4-1 Likelihood function with right truncation

Suppose data are truncated at value k, where k is a positive integer. Probability of a count variable Y with deterministic right censoring is given by

$$p_R(y) = \begin{cases} p(y) & if \quad y < k \\ P(Y \geq k) = \sum_{j=k}^{\infty} p(j) & if \quad y = k \\ 0 & if \quad y > k \end{cases}, \tag{4.1}$$

where the subscript R denotes right truncation. Using this, the expected value of a count variable Y_{CR} in a conditional population with right truncation is obtained as

$$E(Y_{CR}) = \sum_{y=1}^{k-1} \frac{yp(y)}{P(Y>0)} + \sum_{y=k}^{\infty} \frac{kp(y)}{P(Y>0)},$$

which, after some calculation, yields

$$E(Y_{CR}) = k + \frac{1}{1-P(0)} \sum_{y=1}^{k-1} (y-k)p(y). \tag{4.2}$$

Expected value of a count variable Y_{BR} in an on-site distribution with right truncation, on the other hand, is given by

$$E(Y_{BR}) = k + \frac{1}{E(Y)} \sum_{y=1}^{k-1} (y^2 - ky)p(y). \tag{4.3}$$

In general, likelihood function of a count variable with right truncation takes the form

$$L = \prod_{y_i<k} p(y_i) \prod_{y_i=k} P(Y \geq k)$$

from which the log-likelihood function is derived as

$$\log L = \sum_{y_i<k} \log p(y_i) + \sum_{y_i=k} \log P(Y \geq k) \cdot \tag{4.4}$$

In particular, the log-likelihood of a conditional sample with right truncation takes the form

$$\log L_{CR} = \sum_{i=1}^{n} \left[(1-d_i) \log p(y_i) + d_i \log \left(\sum_{j=k}^{\infty} p(j) \right) - \log\{1 - p(0)\} \right] \tag{4.5}$$

where $d_i = \begin{cases} 0 & if \quad y_i < k \\ 1 & if \quad y_i \geq k \end{cases}$.

Likewise, the log-likelihood of an on-site sample with right truncation is given by

$$\log L_{BR} = \sum_{i=1}^{n} \left[(1-d_i) \log y_i p(y_i) + d_i \log \left(\sum_{j=k}^{\infty} j p(j) \right) - \log\{E(Y)\} \right], \tag{4.6}$$

where again, d_i is defined as above and $E(Y)$ denotes the expected value of the count variable Y in the parent population.

4-2 A parametric example

The log-likelihood of a Poisson count data model with right truncation, based on n observations from a conditional sample (x_i, y_i) $i = 1, \ldots, n$ is given by

$$\log L_{CR} = \sum_{i=1}^{n} \left[(1-d_i) \log \frac{e^{-\lambda} \lambda^{y_i}}{y_i!} + d_i \log \left(\sum_{j=k}^{\infty} \frac{e^{-\lambda} \lambda^{j}}{j!} \right) - \log\left(1 - e^{-\lambda}\right) \right],$$

which, after some calculation, yields

$$\log L_{CR} = -n\left\{\lambda + \log\left(1 - e^{-\lambda}\right)\right\} + \sum_{i=1}^{n} \left[(1-d_i) y_i \log \lambda - (1-d_i) \log(y_i!) + d_i \log \left(\sum_{j=k}^{\infty} \frac{\lambda^{j}}{j!} \right) \right].$$

As before, $\lambda = e^{x'\beta}$. Since the estimates cannot be obtained in a closed form, estimation needs to be carried out using numerical methods, for example, the algorithm developed by Berndt, Hall, Hall, and Hausman (1974). The log-likelihood function of an on-site sample with right truncation is obtained in a similar fashion, and is given by

$$\log L_{BR} = -n\log\lambda + \sum_{i=1}^{n}\left[(1-d_i)(\log y_i + y_i\log\lambda - \lambda) + d_i\log\left(\sum_{j=k}^{\infty}\frac{e^{-\lambda}\lambda^j}{(j-1)!}\right)\right].$$

As a more sophisticated example, consider the negative binomial model defined in section 3, that is,

$$p(y\mid x,\lambda,\alpha) = \frac{\Gamma(\alpha+y)}{\Gamma(\alpha)\Gamma(y+1)}\left(\frac{\alpha}{\lambda+\alpha}\right)^{\alpha}\left(\frac{\lambda}{\lambda+\alpha}\right)^{y},$$

allowing for multiplicative heterogeneity. Log-likelihood of this model based on a conditional sample with right truncation is given by

$$\log L_{CR} = \sum_{i=1}^{n}\left[(1-d_i)\left\{\log\frac{\Gamma(\alpha+y_i)}{\Gamma(y_i+1)} + y_i\log\frac{\lambda}{\alpha+\lambda}\right\} + d_i\log\left(\sum_{j=k}^{\infty}\frac{\Gamma(\alpha+j)}{\Gamma(j+1)}\left(\frac{\lambda}{\alpha+\lambda}\right)^{j}\right)\right]$$
$$-n\log\Gamma(\alpha) - n\log\left[1-\left(\frac{\alpha}{\alpha+\lambda}\right)^{\alpha}\right] + n\alpha\log\left(\frac{\alpha}{\alpha+\lambda}\right).$$

For on-site sample with right truncation, the corresponding log-likelihood is derived, after some calculation, as

$$\log L_{BR} = -n\log\lambda - n\log\Gamma(\alpha) + n\alpha\log\left(\frac{\alpha}{\alpha+\lambda}\right)$$
$$+\sum_{i=1}^{n}(1-d_i)\left\{\log y_i + \log\frac{\Gamma(\alpha+y_i)}{\Gamma(y_i+1)} + y_i\log\frac{\lambda}{\alpha+\lambda}\right\} + \sum_{i=!}^{n}d_i\log\left\{\sum_{j=k}^{\infty}\frac{\Gamma(\alpha+j)}{\Gamma(j)}\left(\frac{\lambda}{\alpha+\lambda}\right)^{j}\right\}.$$

Here again, the estimates cannot be obtained in closed form, and numerical methods need to be employed in order to obtain approximate solutions.

5. Concluding remarks

Conditional sampling and on-site sampling are two convenient ways of gathering data on discrete variables. When employing these sampling methods, however, one must be careful which method he or she is using, to correctly assess the type and magnitude of the bias involved. Failure to acknowledge the distributional feature of each sampling scheme results in a misleading conclusion. It must also be emphasized that since the two methods fail to give information on zero values, these methods do not provide any means to test whether or not zero values are generated from the same distribution.

In many cases, it is indeed natural to assume a fundamental difference between a zero and a

non-zero factor. A typical model that accommodates this type of departure is the hurdle model. This model assumes that

$$p(y)=\begin{cases} p_1(0) & if \quad y=0 \\ (1-p_1(0))\dfrac{p_2(y)}{1-p_2(0)} & if \quad y>0 \end{cases},$$

where p_1 and p_2 denote possibly distinct probability functions. It collapses to the regular model when $p_1(y)=p_2(y)$ for every y. Using conditional or on-site sampling method, an estimate for p_2 is obtainable using the arguments in this paper. Another source of information is necessary, however, to estimate p_1 and/or to test whether $p_1(y)=p_2(y)$ for every y.

Often, elaborate models are employed to describe the distributional form of a count variable, while it is simply assumed that data will be gathered through a random sampling of the parent population and the sampling scheme is hardly discussed. This unfortunately, does not necessarily hold in many empirical studies, and when the sampling scheme is itself biased, the obtained estimate will also be biased regardless of how precise the model specification and estimation algorithms are. It is of great importance, therefore, to pay close attention to the sampling scheme, and to employ estimation procedures appropriate for the sampling method adopted.

References

1) Berndt, E. R., Hall, B.R., Hall, R. E., Hausman, J.H. [1974], "Estimation and inference in non-linear structural models", *Annals of Economic and Social Measurement* 3(4), 653-665.
2) Caudill, S. B. and Mixon Jr., F. G. [1995], "Modeling household fertility decisions: estimation and testing of censored regression models for count data", Empirical Economics, 20,183-196.
3) Cameron, A. C. and Trivedi, P. K.[1998], *Regression analysis of count data*, Cambridge University Press.
4) Gurmu, S. [1997], "Semi-parametric estimation of hurdle regression models with an application to medicaid utilization", *Journal of Applied Econometrics*, 12, 225-242.
5) Gurmu, S. and Trivedi, P. K. [1996], "Excess zeros in count models for recreational trips", *Journal of Business & Economic Statistics*, 14,4 69-477.
6) Lambert D. [1992], "Zero-inflated Poisson regression, with an application to defects in manufacturing", *Technometrics*, 34, 1-14.

7) Momma, Maki [2000], "On-site sampling and generalized count data models", *The Economic Review of Toyo University*, 26, 149-168.

8) Mullahy, J. [1997], "Heterogeneity, excess zeros, and the structure of count data models", *Journal of Applied Econometrics*, 12, 337-350.

9) Santos Silva, J. M. C. [1997], "Unobservables in count data models for on-site samples", *Economics Letters,* 54, 217-220.

10) Shaw, D. [1988], "On-site samples' regression --problems of non-negative integers, truncation, and endogenous stratification", *Journal of Econometrics*, 37, 211-223.

11) Winkelmann, R. [1997], *Count Data Models:Econometric Theory and Application to Labor Mobility*, Berlin, Springer-Verlag.

12) Winkelmann, R. and Zimmermann, K. F. [1995], "Recent developments in count data modeling: theory and application" *Journal of Economic Surveys*, 9, 1-24.